Unlocking Scientific Concepts: How Effective Are LLM-Generated Analogies for Student Understanding and Classroom Practice?

Zekai Shao*
Fudan University
Shanghai, China
gemini25szk@gmail.com

Siyu Yuan*
Fudan University
Shanghai, China
syyuan21@m.fudan.edu.cn

Lin Gao
Fudan University
Shanghai, China
lgao.lynne@gmail.com

Yixuan He
Fudan University
Shanghai, China
1265188674@qq.com

Deqing Yang†
Fudan University
Shanghai, China
yangdeqing@fudan.edu.cn

Siming Chen†
Fudan University
Shanghai, China
simingchen3@gmail.com

Abstract

Teaching scientific concepts is essential but challenging, and analogies help students connect new concepts to familiar ideas. Advancements in large language models (LLMs) enable generating analogies, yet their effectiveness in education remains underexplored. In this paper, we first conducted a two-stage study involving high school students and teachers to assess the effectiveness of LLM-generated analogies in biology and physics through a controlled in-class test and a classroom field study. Test results suggested that LLM-generated analogies could enhance student understanding particularly in biology, but require teachers' guidance to prevent overreliance and overconfidence. Classroom experiments suggested that teachers could refine LLM-generated analogies to their satisfaction and inspire new analogies from generated ones, encouraged by positive classroom feedback and homework performance boosts. Based on findings, we developed and evaluated a practical system to help teachers generate and refine teaching analogies. We discussed future directions for developing and evaluating LLM-supported teaching and learning by analogy.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; User interface toolkits; Field studies;

Keywords

Analogy Generation, Large Language Models, Scientific Concept Understanding, Classroom Study

ACM Reference Format:
Zekai Shao, Siyu Yuan, Lin Gao, Yixuan He, Deqing Yang, and Siming Chen. 2025. Unlocking Scientific Concepts: How Effective Are LLM-Generated Analogies for Student Understanding and Classroom Practice?. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3714313

*Zekai Shao and Siyu Yuan contributed equally to this research.
†Siming Chen and Deqing Yang are the corresponding authors.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714313

1 Introduction

Analogy facilitates the comprehension of complex concepts by associating them with familiar ones [12, 21, 23, 50]. It plays a crucial role across various domains, particularly in education, science and problem-solving [30, 53, 62, 72, 74, 76]. It enhances cognitive processes such as creativity [29, 33, 39], aids in effective communication [6, 18], and facilitates the learning and understanding of complex concepts. Analogies, such as "water waves" for "light waves", "the solar system" for "atomic structure", and "hydraulic pump" for "heart" are frequently employed in textbooks and classroom teaching to assist students in understanding scientific concepts. Analogies can boost student understanding of concepts by opening new perspectives, making abstract ideas more relatable by connecting them to familiar situations, assisting in visualizing these concepts, sparking students' interest and motivation, and taking into account their previous knowledge to reveal any misconceptions [16].

The rapid advancement of Large Language Models (LLMs) has led researchers to employ these models in generating analogies that enhance concept comprehension [54]. Unlike smaller language models (LMs), such as BERT [13] and GPT-2 [59], that primarily address word-pairing analogies (e.g., "king is to man as queen is to woman") [4, 8, 28, 49], LLMs are capable of creating more complex, free-form natural language analogies [2]. In specific use cases, for example, researchers have explored using LLMs to generate biologically inspired analogies to foster scientific ideation [37] and to transform abstract data into vivid data analogies that enhance the understanding of readers [10].

Initial research has explored the use of LLMs to generate concept-related analogies to assist students and teachers [3]. However, the effectiveness of LLM-generated analogies in educational settings remains underexplored, highlighting the need for a thorough evaluation to guide teachers and students in their practical application. We draw on two traditional education scenarios on evaluating human-made analogies to assess LLM-generated analogies: students solving problems using analogies without human intervention [5, 25, 26, 72], and teachers employing analogies in the classroom [53, 61, 76]. Evaluating LLM-generated analogies in the

first scenario determine their effectiveness in assisting students with problem-solving by accuracy. It also offer insights for developing LLM-assisted self-learning tools [19, 48] that produce more beneficial educational analogies. In the second scenario, evaluation helps to understand the needs and practical effectiveness of LLM-generated analogies in real classroom practices with teachers' instruction. It also provides insights of needs for developing LLM-assisted tools to help teachers teach with analogies. Therefore, we aim first to evaluate the effectiveness of LLM-generated educational analogies in helping students grasp scientific concepts across two distinct educational settings. Based on the findings, we then consider developing a practical system to explore the practical application of LLMs for analogy generation in supporting teachers and students.

Evaluating LLM-generated analogies in practical applications is challenging, requiring manual annotation [68] and human-subject study involving diverse participants across varied settings. Previous studies have evaluated the use of LLM-generated analogies in creative tasks, such as design problem reformulation [15]. However, due to the varied participants and cognitive demands of education, the methodologies and findings from these domains are not transferable to educational settings. To address the gap, we answered the following research questions (**RQ**s):

- **RQ1:** How effective are LLM-generated analogies for student understanding without human intervention?
- **RQ2**: What kind of analogies do teachers need LLMs to generate for classroom practice?
- **RQ3**: How effective are these LLM-generated analogies for classroom practice with teacher intervention?

To answer these **RQ**s, we design a two-stage study to evaluate the effectiveness of LLM-generated analogies in helping students understand scientific concepts in educational settings. We use GPT-4o [54], the state-of-the-art LLM, to generate analogies for concepts in physics and biology, and conduct human-subject study in a Chinese high school. We addressed **RQ1** through a controlled in-class students test in **Study I**: the experimental group used both textbook explanations and carefully generated analogies to answer questions about unknown concepts, while the control group relied only on textbook explanations. The effectiveness of LLM-generated analogies was primarily measured by accuracy, with self-confidence rating evaluating students' subjective satisfaction. Interviews with participating students provided additional insights. In **Study II**, we conducted pre-class interviews with teachers and senior students to understand the need for LLM-generated analogies in classroom teaching (**RQ2**). Interview findings helped us design the following field study and refine analogy generation. We then carried out a one-week field study with twelve lessons in which teachers selected and modified the generated analogies based on their needs, using them in one class while maintaining regular teaching in the other. Through observations and teacher interviews, we qualitatively evaluated the effectiveness of LLM-generated analogies for classroom teaching with teacher intervention (**RQ3**).

Based on the findings from both studies, we designed an interactive system that uses LLM to assist teachers in preparing analogies. We first interviewed two teachers to assess what they want from the system. Using insights from the interviews and studies, we defined the design requirements to guide system development. We invited six physics and biology teachers from different schools to participate in the system evaluation. After tutorial and free exploration to confirm their familiarity with the system, teachers used the system to create analogy of scientific concepts for teaching over a week. Based on users' data records and interviews, we demonstrated the practical effectiveness of LLMs in supporting teaching by analogy and then discussed future research directions.

The main **contributions** are summarized as follows:

- We are among the first to design and conduct a two-stage study, comprising a controlled experiment and a field study, to evaluate the effectiveness of LLM-generated analogies in student understanding and classroom practice.
- We contribute empirical evidence and new knowledge into educational LLM-generated analogies, revealing that their effectiveness without intervention varies on subject characteristics and can lead to student overconfidence, while in classroom practice, analogies are refined by teachers to align with their teaching focus and preference, enhancing both classroom and homework performance and inspiring new teaching methods.
- Based on empirical evidence and new knowledge, we developed a practical system to help teachers build analogies of scientific concepts, conducted a system evaluation demonstrating its effectiveness, and provided design implications for future development and evaluation of LLM-assisted education with analogy.

2 Related Work

This section reviews related work on analogy in education, evaluating analogy in HCI, analogy-making with LLMs, and LLM-assisted educational systems.

2.1 Analogy in Education

Analogies help humans understand complex concepts by linking them to familiar ones, making them a valuable tool in educational contexts. Many studies [5, 25, 26, 72] have investigated analogical problem solving, where students of various ages solve unfamiliar problems using well-designed analogies. Through observational feedback and statistical analysis, researchers have established frameworks and several guidelines for using analogies in education. For example, as discussed by [25], the source of the analogy would share similar relationships with the target, yet originate from a semantically distant field. However, such lab studies often involve experimenters posing problems, with students merely solving them without instructional guidance [5], which diverges from real classroom learning. Therefore, further research [53, 61, 74, 76] have investigated how teachers and students engage with analogies in classroom settings, leading to nuanced insights on the influence of students' age and background and teachers' strategies.

Although previous studies have explored the characteristics and use of analogies in education, they have not examined those generated by LLMs, which is crucial given the growing importance of LLM-assisted education [19, 48]. Our work fills this gap by leveraging LLMs to generate analogies tailored to specific education needs, incorporating established characteristics from prior literature and our interviews. We design human-subjective studies to

evaluate their effectiveness in problem-solving tests and classroom environments following prior research.

2.2 Evaluating Analogy in Human-Computer Interaction

Analogy has long been studied in HCI for its effectiveness in various context, including algorithms improvement [1, 57], cancer communication [32], narrative framing [77], enhancing deliberation [81], communicating standardized effect sizes [42], and sensemaking of LLM responses [24].

Two key research directions about analogies in HCI are for enhancing numerical comprehension through data analogy and fostering creativity. Data analogies link abstract data to familiar concepts to improve understanding. Researchers evaluate these analogies using controlled experiments and assess effectiveness through subjective ratings like helpfulness [10, 35, 43, 64, 66], estimation errors [35, 64], and correlations between model and human ratings [66]. Analogies also facilitate scientific discovery and design. In scientific discovery, evaluations involve coding analogy types [7, 38], calculating similarity metrics [7], and conducting think-aloud sessions with scientists [38]. For creative design, analogies are assessed by novelty [9, 83, 88], quality [9, 82], relevance and domain distance [27], feasibility [88], and rationality [9]. Recently, Ding et al. [15] explored GPT-3's capacity to augment cross-domain analogical reasoning, finding it helpful for creative problem reformulation despite the risks of harmful content.

However, there has been limited exploration of analogy search in HCI for education [44]. While researchers have adopted LLMs to help students and teachers generate novel analogies [3], systematic evaluations of their effectiveness in educational settings are lacking. Given the unique cognitive demands of education, existing assessments [15] may not be directly applicable. Our work aims to address this gap and offer insights into analogy generation for education.

2.3 Analogy-making with Language Models

Analogy is vital for human cognition and has attracted considerable interest from the AI research community. Traditionally, studies on analogy-making in AI have concentrated on creating word analogies (e.g., "king is to man as queen is to woman") using smaller language models (LMs), e.g., BERT [14] and GPT-2 [60] trained on specific datasets [4, 8, 28, 49, 75, 86]. With the advancement of LLMs [54, 56, 71, 73], there has been a shift toward generating natural language analogies, i.e., free-form analogies [2, 15, 34, 36, 67, 78, 79] and forming structural analogies [68, 85]. Researchers typically design prompts manually for free-form analogies to guide LLMs in generating analogies [2, 78]. For example, Bhavya et al. [2] constructed a new dataset including standard science analogies and science analogies from academics and adopted prompt engineering to ask LLMs to generate analogies. The results show that LLMs are sensitive to prompt design, temperature, and injected spelling errors, particularly the distinction between questions and imperative statements. We followed their optimal prompt format for our generation process. For evaluation of the generation quality of analogy, previous studies have relied on annotators manually evaluating analogies according to established principles of analogy cognition [68].

In contrast to these approaches, our study is pioneering in investigating how analogies generated by LLMs can help students understand scientific concepts. We analyze the characteristics of analogies in educational settings through literature reviews and interviews and incorporate them into prompts for generation. Then, we use LLMs to generate educational analogies and evaluate them in real tests, classroom practice, and a practical system.

2.4 LLM-assisted Educational Systems

With the rapid advancement of LLMs, researchers are exploring their potential to develop efficient and practical systems that support students and teachers in educational tasks [40, 80]. For students, many studies have focused on creating intelligent tutoring systems powered by LLMs. Examples include enabling fully autonomous self-learning pipelines to support self-regulated learning [19] and developing and evaluating LLM-based learning assistants in classroom settings [41, 48]. For teachers, several LLM-based systems are designed to effectively monitor and analyze students' learning activities [51, 70, 87]. In addition, researchers aim to assist teachers in creating diverse teaching materials, such as lesson plans [17], diagrammatic problems [52], and reading quizzes [47].

Our work explores a novel aspect of LLM-driven education: evaluating the effectiveness of LLMs in generating teaching analogies. One preliminary research has initially explored generating educational analogies [3], while its system design lacks the support of empirical evidences and fails to address teachers' needs. Instead, we first conducted a two-stage study to gain insights and empirical evidence and identify needs for teachers and students. We then developed and tested a system to support teachers in creating and refining analogies for lesson preparation and discussed future integration with diverse LLM-based educational tools for various users.

3 Method

In this section, we introduce the overview, our study design, and the techniques for analogy generation with LLM.

3.1 Overview

Our work aims to first understand LLM-generated educational analogies' effectiveness through empirical studies and then design practical LLM-assisted educational systems leveraging findings from studies. For empirical studies, we explore two study settings: one where students solve problems using only LLM-generated analogies and necessary materials without additional guidance (Sec. 4), and another where teachers integrate LLM-generated analogies flexibly into classroom instruction (Sec. 5), considering the following two reasons. First, these two settings align with those used to evaluate human-made analogies in traditional education research: student-only testing [25, 72] and teacher-led classroom practice [53, 76]. Second, evaluating in two settings respectively inform the design of systems that incorporate LLM in generating analogies to (1) support self-learning for students [19] and (2) boost teaching for teachers [17], within the context of LLM-assisted education research.

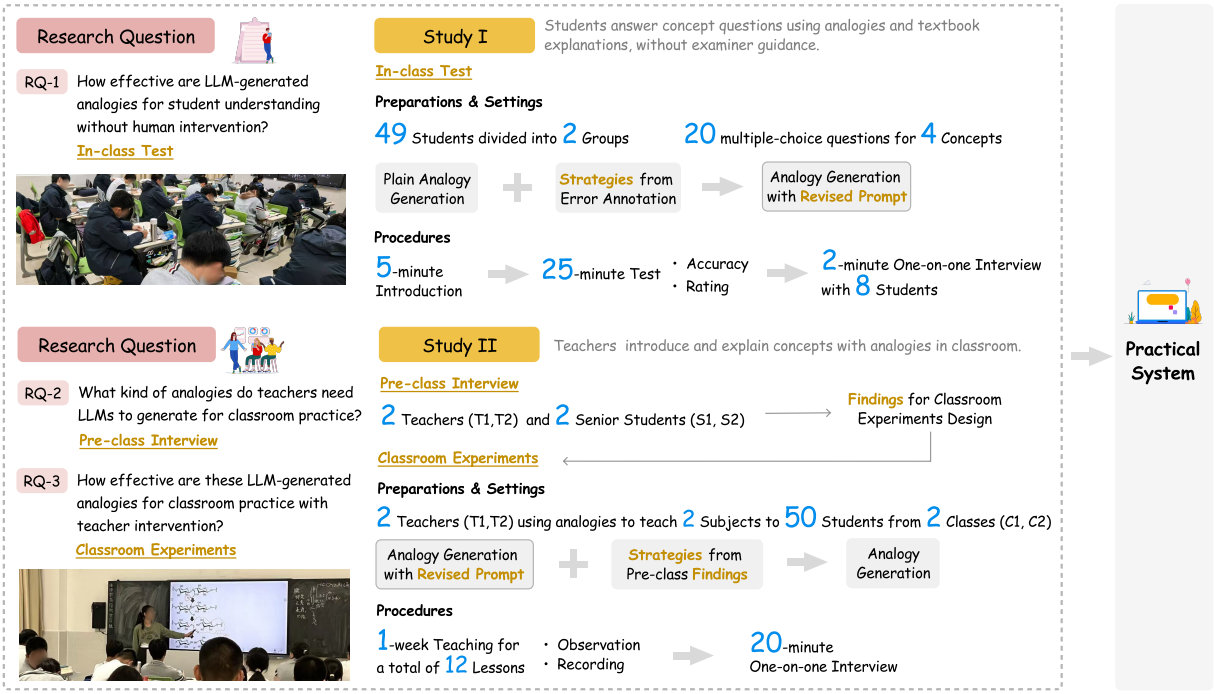

Figure 1: Our Study I address RQ1 through an in-class test. Study II address RQ2 through a pre-class interview and RQ3 through a one-week classroom study. Studies' findings lead us to build an LLM-supported practical system for analogical education.

Based on study findings, we identify a more feasible direction between supporting student self-learning and teacher instruction by analogy for designing a practical system (Sec. 6). Drawing from study findings and system evaluation, we offer implications for future LLM-assisted educational systems with analogy (Sec. 7).

3.2 Study Design

For both studies, we need to ground the effectiveness of LLM-generated analogies on improving students' concept understanding by comparing the accuracy of problem-solving across different student groups. Specifically, we may need to compare students' problem-solving accuracy in classroom teaching with and without LLM-generated analogies. This comparison is driven by the unique features of LLM-generated analogies compared to traditional ones, as well as teachers' potential unfamiliarity with them and unclear expectations. Unlike classic analogies refined and validated over generations, they may be harder for teachers to adapt to support student understanding. We will confirm this design in the interview with teachers (Sec. 5.1.3). Besides accuracy, it's important to evaluate students' and teachers' subjective satisfaction [48], reflected through subjective ratings, classroom feedback, and interviews.

As shown in Fig. 1, we conducted a two-stage study in a Chinese high school on LLM-generated analogies for physics and biology concepts to explore this topic.

3.2.1 Study I. We conducted an in-class test to evaluate the effectiveness of LLM-generated analogies in understanding concepts without human intervention (**RQ1**).

- **Participants.** 49 Chinese high school freshmen from 2 classes.
- **Procedure.** Students were divided into two groups: one received LLM-generated analogies, while the other did not. They then completed an in-class test with 20 multiple-choice questions for 4 concepts they didn't learn. Afterward, 8 students participated in interviews.
- **Measure.** Effectiveness was assessed through quantitative results of students' answer accuracy and confidence ratings, and qualitative insights from student interviews.

3.2.2 Study II. Study II consists of two sub-studies. We conducted a pre-class interview to identify classroom needs for LLM-generated analogies (**RQ2**).

- **Participants.** 2 Chinese teachers and 2 Chinese senior students.
- **Procedure.** The interview followed a semi-structured format, allowing participants to discuss their experiences, expectations, and concerns on using LLM-generated analogies in the classroom.
- **Measure.** We identified qualitative findings on their perceptions of analogy use during the interview.

We then conducted a controlled field study to evaluate the effectiveness of LLM-generated analogies in classroom practice (**RQ3**).

- **Participants.** The 2 teachers from the pre-class interview and 50 Chinese students from 2 classes.
- **Procedure.** The teachers taught both classes over one week, delivering a total of 12 lessons. In one class, they incorporated LLM-generated analogies, while in the other, they followed regular instruction without analogies as a control.
- **Measure.** We derived qualitative findings from teachers' selection and modification of analogies and student feedback.

3.3 Techniques for Analogy Generation with LLM

With the development of LLMs, their capabilities have demonstrated significant potential in generating satisfying content. Therefore, recent studies have utilized LLMs to create analogies using manually-designed instructions. We align with this approach by crafting prompts for LLMs grounded in established analogy theories and our interviews. The prompt consists of three parts:

- **Task Description** demonstrates the task that LLMs need to achieve.
- **Principles** highlight the requirements, rules, and constraints that LLMs must follow to complete the task.
- **Input Resource** lists the input materials needed to complete the task.

Given the lack of standardized guidelines for using LLMs in educational analogy generation, we summarize principles from educational literature [20, 22, 31] and refine them through manual annotation in **Study I** and interviews in **Study II**. Additionally, inspired by prior work [84], Study I uses an over-generation and filtering strategy to select the best analogies, while Study II leaves all candidates for teachers to refine. The techniques used in the further developed system is determined by the results of two studies.

4 Study I

In this section, we detail the participants, data preparation process, stimuli, procedure, and results analysis of Study I.

4.1 Participants

Two classes of freshmen, totaling 49 Chinese students from a Chinese high school with which we have a scientific research collaboration, participated in Study I. Their ages ranged from 15 to 17, with 26 males and 23 females. They had recently started high school physics and biology courses. Their entrance exam scores and classroom performance suggested a normal cognitive level, and we did not pre-select students based on their abilities.

4.2 Data Preparation

As shown in Fig. 2, we began by manually selecting ten scientific concepts from physics and biology in Chinese high school textbooks. Next, we used the advanced LLM, GPT-4o [54] (temperature = 0.7) to generate three analogies for each concept with three principles summarized from education research [20, 22, 31] incorporated in the prompt. Three authors independently identified and annotated errors in generated analogies. After repeated discussion, the annotators classified the errors into four types: two related to factuality and two to consistency, as follows.

- **Analogy Object Paradox**: The objects of the analogy do not align with physical laws or commonsense knowledge.
- **Inappropriate Analogy**: The analogy fails to accurately mirror the concept, leading to misconceptions.
- **Object Confusion**: The same analogy objects are assigned different roles or functions across various contexts.
- **Logical Contradiction**: The syntax within a sentence or paragraph contradicts itself.

The inter-rater reliability among annotators reached Fleiss' Kappa of 0.83 for Analogy Object Paradox, 0.94 for Object Confusion, and 1 for the remaining error codes. Error annotations in subsequent steps achieved similar reliability. As shown in the first row of Tab. 2, out of the 30 generated analogies, 16 were correct. The remaining analogies frequently exhibited the first three error types with one analogy containing logical contradiction. From these errors, we derived four new principles and added them to the prompt template (Tab. 1 I) to help GPT-4o avoid these errors. However, even with these improvements, GPT-4o still made errors. To address this, we followed prior AI research [46, 58, 84] and allowed GPT-4o to automatically select the best of the three candidate analogies generated for each concept. The prompt for analogy selection is shown in Tab. 1 II. As shown in the third row of Tab. 2, enabling the model to self-correct improved the accuracy of the analogies.

Finally, we further categorized ten analogies selected by the LLM into four distinct groups, as illustrated in Fig. 3:

- **Correct and Satisfying Analogy**: Analogies in this category are error-free. The objects in these analogies are realistic, align with common sense, and adhere to physical laws, effectively and vividly illustrating scientific concepts.
- **Correct Analogy with Imagination**: Analogies in this category require envisioning non-existent objects or processes to explain a concept. While logically sound, they demand creative thinking and imagination from students.
- **Correct Non-Analogy**: This is more akin to an example-based explanation than a true analogy and is not generally recognized as an analogy in cognitive science.
- **Incorrect Analogy**: This category includes analogies exhibiting previously identified error types. These analogies are inappropriate for students to refer to, as they do not accurately convey the intended concept.

The analogies for the five biological concepts fell under the **Correct and Satisfying Analogy** category. In contrast, the five physical concepts were distributed as follows: three under **Incorrect Analogy**, one under **Correct Analogy with Imagination**, and one under **Correct Non-Analogy**.

We limited the number of concepts and analogies to four to avoid overwhelming students with too much new knowledge in further tests. To ensure a balance across subjects and categories, we specifically selected two biological concepts categorized as **Correct and Satisfying Analogy** and two physical concepts, one each under **Correct Analogy with Imagination** and **Incorrect Analogy**, as shown in the right side of Fig. 3. We excluded the one **Correct Non-Analogy** from further consideration, as it is not typically classified as an analogy.

4.3 Stimuli

Since students do not have access to electronic devices and are more familiar with and serious about traditional classroom tests, we conducted offline tests in class.

Based on the data preparation, we printed a test paper and two reference materials. The test paper comprises 20 multiple-choice questions, with 5 questions assigned to each of the following 4 concepts: Nuclear Fission and Fusion, Wave-Particle Duality, Blood

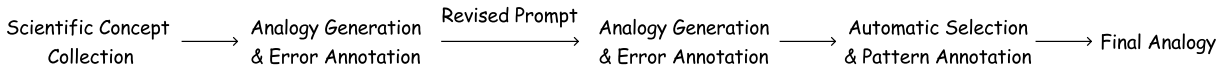

Figure 2: The pipeline for analogy generation to explain scientific concepts for data preparation in Study I.

Table 1: Prompt template for GPT-4 to generate analogies (I) and select the analogy from three candidates for the given concept (II). *Green texts* are new principle after revising.

I: Analogy Generation with Revised Prompt
/ Task Description */* Your task is to use an analogy to explain the scientific concept to students. Here are some principles for generating appropriate analogies: */* Principles */* 1. The similarity between the objects in the analogy and those in the scientific concept should be minimal. 2. The relationships in the analogy and the scientific concept should be highly similar. 3. The analogy should use objects that students are very familiar with from everyday experiences. *4. The analogy should accurately identify similar relationships with the scientific concept and avoid forcing non-existent similarities.* *5. The objects in the analogy and the scientific concept should align with scientific laws and commonsense knowledge.* *6. An object in the analogy cannot have different roles or functions in different contexts.* *7. The logic within a sentence or paragraph should not be self-contradictory.* */* Input Resource */* To generate the appropriate analogy according to students' learning progress, we provide you with textbook content related to this scientific concept for your reference. The textbook content: `{input_text}` The scientific concept: `{concept}` Analogy:
II: Analogy Selection
/ Task Description */* There are three candidate analogies which are used to explain the scientific concept based on textbook content. Your task is to select the best analogy from these three candidates. Here are some principles for generating appropriate analogies: */* Principles */* Same as principles in I (Omit) */* Input Resource */* The textbook content: `{input_text}` The scientific concept: `{concept}` The generated analogies: Candidate 1: `{analogy_1}` Candidate 2: `{analogy_2}` Candidate 3: `{analogy_3}` You need to give reasons first and then give the answer with the format: `Final Answer: Candidate X` Answer:

Table 2: Errors and accuracy of Plain Generation (Plain), Revised Generation (Revised), and Automatic selection ($\text{Selection}_{\texttt{Auto}}$). (Data #) shows the number of data. (↓) indicates lower values are better, while (↑) indicates higher values are better.

Process	**Data #**	**Factuality**		**Consistency**		**Accuracy (↑)**
		Analogy Object Paradox (↓)	**Inappropriate Analogy (↓)**	**Object Confusion (↓)**	**Logical Contradiction (↓)**	
Plain	30	**0.27**	0.23	0.23	0.03	0.53
Revised	30	0.33	0.23	**0.17**	**0.00**	0.53
$\text{Selection}_{\texttt{Auto}}$	10	0.30	**0.20**	0.20	**0.00**	**0.60**

Sugar Regulation, and Immune Response. In addition to selecting answers, students are required to complete a 5-point Likert scale rating for self-confidence to measure their subjective satisfaction. The first reference material provides textbook explanations for the four concepts, while the second adds LLM-generated analogies before the explanations. The test paper and the reference materials present the concepts in the same order. They are highlighted in bold, making it easier for students to find and connect the information with the questions.

4.4 Procedure

Then, we conducted an in-class test for students in two classes lasting 30 minutes. We first gave a 5-minute introduction for the background of our test. After the introduction, we randomly divided the students into two groups and distributed two sets of reference materials to each group. We clarified the meaning of self-confidence rating. Under our supervision, each student then independently completed the test using the materials provided in 25 minutes.

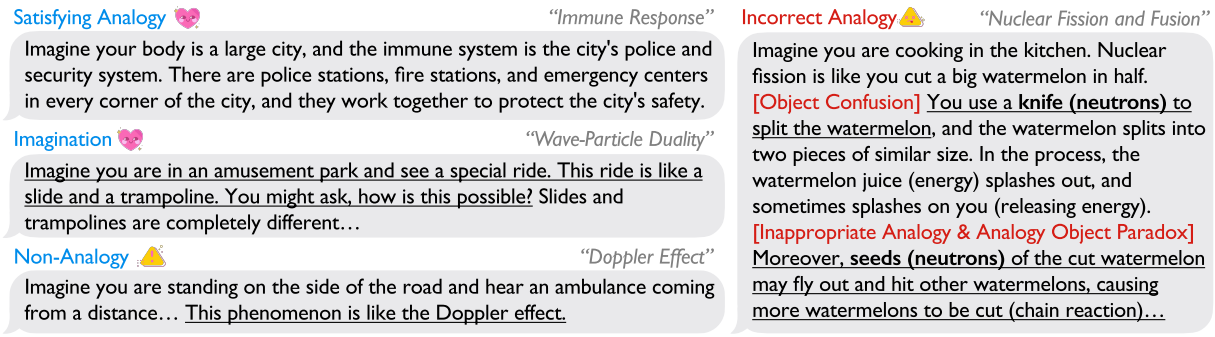

Figure 3: The examples of final analogies generated from GPT-4o after iterative generation and annotation in Study I.

After the test, we interviewed four students from each group, totaling eight participants. Each 2-minute interview earned participants a $2 gift card. We asked them about any difficulties during the test and, for those with analogies in their materials, how these helped them answer questions alongside textbook concepts.

4.5 Results Analysis

Our selection criteria excluded test papers with incomplete answers. After examining the 49 test papers, we excluded 5 that had more than 5 unanswered questions. The remaining 44 fully completed papers, 22 from each group, were considered valid data. Our analysis followed a top-down approach, starting from the overall test (20 questions) and proceeding to finer levels: subject (10 questions each), concept (5 questions each), and individual questions. For further comparison, students' responses were averaged across questions at the first three levels.

We computed descriptive statistics to gain overall insights and performed statistical tests to determine significance at each level. The experiment results include the students' answer accuracy and confidence ratings, both can be regarded as ordinal categorical variables. Thus, we mostly employed the exact Wilcoxon-Mann-Whitney test (using the R package `coin`) to evaluate the significance of difference between the two groups. For one exception, we employed Fisher's exact test (using the R package `stats`) on students' answer accuracy at the individual question level, where the accuracy is binary (either 0 or 1). In any of the tests, a small p-value indicates a potential association between the use of LLM generated analogies and the students' outcomes, and a significance level is defined as p<0.05 in all tests. We also calculated Kendall's tau correlation coefficient to assess the relationship between students' objective answer accuracy and subjective confidence ratings within each concept and group.

We summarize our findings as follows. The two groups are referred to as Group T (Textbook explanation only) and Group L (Textbook explanation with LLM-generated analogy), while the interviewed students are denoted as T1-T4 and L1-L4.

LLM-generated analogies generally aid problem-solving and have a greater impact on biological concepts than physical concepts. As shown in the left of Fig. 4, the overall accuracy for physics questions is higher, while a significant association exists between accuracy and group for biology questions (p = 0.042).

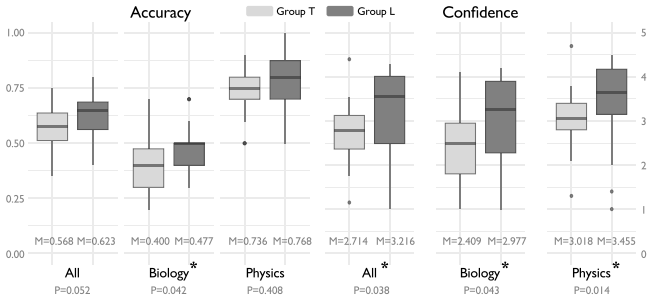

Figure 4: Boxplots showing the distribution of answer accuracy and confidence ratings for the two groups and comparing the overall test results and the two subjects. "M" represents Mean, "p" represents the significance of the association between accuracy and group, determined by the exact Wilcoxon-Mann-Whitney test, and * represents significance (p<0.05).

Examining individual questions (Fig. 5), there are three questions within the two biological concepts where Group L's accuracy largely exceeds Group T's by more than 0.25. Besides, a strict Fisher's exact test shows marginally significant associations between accuracy and group for two questions (Q1 and Q3 of "Immune Response") (p = 0.067). However, no such clear difference is seen for the physics questions. In the interviews, all four students from Group L (L1-4) coincidentally explained the role of analogies based on subjects. They noted that explanations for physical concepts are relatively concise, allowing them to understand directly without analogies. In contrast, the lengthy explanations for biological concepts made analogies helpful to "*get an overview and quickly identify key terms*", as indicated by L4.

LLM-generated analogies may negatively affect students' understanding without teacher intervention due to errors and missing information in analogies and students' incorrect learning strategies with over-reliance. Although some students in Group L identified the analogy of "Nuclear Fission and Fusion" as an **Incorrect Analogy** and noted specific LLMs' hallucination during the interview, Group L's accuracy on all five questions was no higher than Group T's (Fig. 5). Furthermore, we also found that, although the **Correct and Satisfying Analogies** slightly

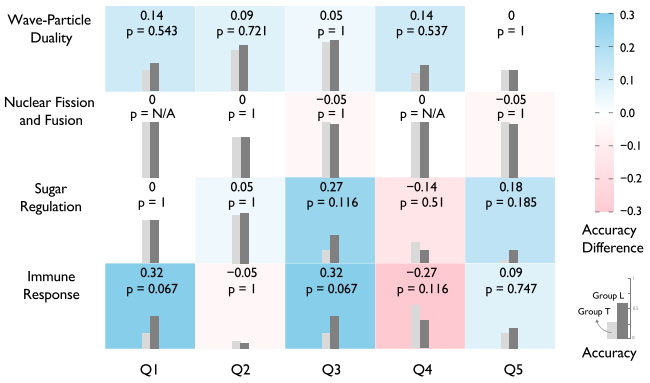

Figure 5: Heatmap of accuracy differences between Group L and Group T for individual questions, with blue indicating higher accuracy for Group L and red for Group T. Each cell contains a bar chart of the respective accuracies and a p-value representing the significance of the association between accuracy and group, determined by Fisher's exact test. For two of the questions, since all students in both groups answered them correctly, Fisher's exact test is not applicable, and p = N/A.

improved overall answer accuracy, they could also harm students' understanding in some cases due to missing information. For example, Group L's accuracy for Q4 of "Immune Response" was 0.27 lower than Group T's. Their incorrect answer choices suggest that some students believed that "plasma cells can recognize antigens." However, the textbook explains that "antigen-presenting cells such as B cells recognize antigens" and "plasma cells release antibodies to eliminate antigens", while the LLM-generated analogy only includes the latter and omits the former. This may be linked to over-reliance issue, as L1 and L3 described their learning strategies during the interview as "*reading the analogy first, answering the questions, and not revisiting the textbook if the answers seemed clear from the analogy.*"

Students subjectively appreciate the correct LLM-generated analogies often with overconfidence. We were surprised to find the strongest association between group and self-confidence was for "Wave-Particle Duality" ($p = 0.025$) among the four concepts. This suggests that students were receptive to the **Correct Analogy with Imagination**. However, there was no significant association between group and answer accuracy for this concept. We also observed overconfidence among **Correct and Satisfying Analogies** for biological concepts: a significant association between group and confidence ratings for Q2 of "Blood Sugar Regulation" ($p = 0.034$), but none with accuracy ($p = 1$). As shown on the right of Fig. 4, there are significant associations between group and confidence ratings for both subjects. Besides, We found a negligible correlation between accuracy and the confidence rating, as the absolute value of Kendall's tau correlation coefficient between them in each group and concept was ≤ 0.2.

Overall, our empirical evidences suggest that LLM-generated analogies are currently unsuitable for unsupervised self-learning systems. We will discuss LLMs in supporting students' learning by analogy in Sec. 7.4.

5 Study II

In this section, we introduce the pre-class interview (Sec. 5.1) to gather requirements for the classroom experiments (Sec. 5.2) that evaluate the actual use of LLM-generated analogies in classroom teaching.

5.1 Pre-class Interview

In this subsection, we outline the participants for our pre-class interview (Sec. 5.1.1), the procedure and stimulus (Sec. 5.1.2), and the findings and derived requirements (Sec. 5.1.3) for further classroom experiments.

5.1.1 Participants. We recruited two Chinese teachers and two Chinese students from the same high school as Study I to participate in pre-class interviews. The two teachers (T1 and T2; 1 female) are a physics teacher with 6 years of teaching experience and a biology teacher with 3 years of teaching experience. Both have a bachelor's degree, are interested in AI-assisted education, and teach first-year high school courses in the semester during our study. The two senior students (S1 and S2; 1 female) are in their third year of high school, have learned the concepts used in Study I, and have above-average grades.

5.1.2 Procedure. We conducted one-on-one semi-structured online interviews with the participants via Tencent Meeting. The student interviews lasted 40 minutes, with a $10 gift card for each student, while the teacher interviews lasted 60 minutes, with a $20 gift card for each teacher.

For teachers, the interview included four steps to understand the requirements of LLM-generated analogies in teaching.

Step 1: Analogy Orientation. We first presented them with common analogies in the classroom from the literature (e.g., "light waves and water waves", "heart and hydraulic pump") [53] to orient them to analogies and ensure the terminology used during the interview. We asked the teachers to recall the analogies they had used and encouraged them to think aloud about any experiences with analogies throughout the interview.

Step 2: Analogy Usage Exploration. After that, we conducted interviews using a questionnaire primarily based on the one proposed by [53] but modified to incorporate insights from the latest research over the past decade. We first investigated how teachers prepare analogies, such as whether they prepare in advance or improvise and adjust in the class. Then we asked teachers to share the characteristics of good analogies in their opinion [30] and whether they agree with the principles we summarized from the literature in Study I. We also investigated whether teachers involved students in building analogies during teaching, at which step of introducing knowledge points they used analogies, and whether they followed the six-step theoretical model about analogy [62]. Other questions covered include whether visual aids were used and whether immediate feedback was provided on students' understanding of analogies.

Step 3: AI Usage Exploration. We then asked teachers about their previous experiences with AI tools like ChatGPT, their familiarity with AI-assisted teaching or self-learning. We also inquired about their concerns on AI performance and its application in educational settings [11, 69], and whether they believe AI could partially

replace teachers to achieve educational goals such as mastering basic concepts, problem-solving, and developing higher-order independent thinking skills.

Step 4: Expectations Sharing on LLM-generated Analogies. We then showed each teacher the analogies from Study I in their respective teaching subjects. We asked them to evaluate each analogy's strengths, weaknesses, classroom applicability, and potential for teacher modification Building on this, we asked teachers to share their expectations for effective AI-generated analogies.

The interviews with the students were focused on their classroom experiences rather than exploring AI usage since their role in the classroom mainly involved receiving information rather than designing analogies, and they spent most of their time without any smart devices or AI. Initially, students were asked to recall analogies used in class. We then presented the ten concepts from Study I, asking students to reflect on their learning experiences. Next, we presented the ten analogies from Study I and asked for their feedback on their effectiveness in enhancing their understanding.

5.1.3 Findings and Derived Requirements. The findings of the teacher interview were summarized in Tab. 3. Based on these findings and the interview with senior students, we conclude the requirements for data preparation and classroom study design as follows.

Providing analogies to teachers during lesson preparation. At the beginning of the interview, both teachers clearly stated that they often use analogies in class, with most being prepared in advance. The physics teacher (T1) frequently referred to analogies found in teaching aids. The biology teacher (T2) listed key knowledge points during lesson preparation and then considered suitable analogies, drawing on personal experience and input from other veteran teachers. Both teachers and students claimed that students rarely participate in the construction of analogies except in student-led discussion sessions.

Generating analogies based on subjects' characteristics and analogy needs. Two teachers demonstrated apparent differences in their need for and use of analogies. Based on their explanations, we attribute these differences to their subjects' characteristics rather than personal preferences. T1 frequently used analogies between concepts like "electric field and magnetic field," noting the abstract nature of physics and the difficulty of finding everyday analogies. In contrast, T2 primarily employed interesting everyday life analogies, such as likening "chromosome crossing over" to "swapping legs between classmates". However, when presented with the analogy for "blood sugar regulation" generated in Study I, T2 suggested it could be analogized with "thyroid hormone regulation", as functions are related and thus easy for students to grasp.

Generating analogies for teaching key points and helping students focus. Both teachers stated that the primary goal of using analogies was to help students understand key concepts. Additionally, they emphasized that some analogies helped students maintain engagement. T1 mentioned, *"When I notice students getting sleepy, I occasionally improvise an interesting analogy related to the concept to wake them up."* T2 used images of people and mummies to explain the dry and fresh weight of cells, which are vivid and engaging without distracting students.

Generating necessary analogies determined by teachers. Both teachers acknowledged our generated vivid analogies in Study I. However, they criticized many of them as being overly complicated and unnecessary. For "nuclear fission and fusion" and "auxin," T1 and T2 pointed out that students could quickly understand them through pictures and animations. Additionally, T1 mentioned that the "molecular kinetic theory" is relatively simple and not a key focus of exams, thus only requiring memorization. T1 also stated that concepts in atomic physics, such as the "photoelectric effect," are too isolated from other physical concepts to be conveyed through analogy. Additionally, students interviewed could not recall many concepts taught using analogies and viewed many analogies in Study I as redundant.

Generating non-complex analogies for certain aspects of the concept. Both teachers emphasized the importance of analogizing only parts of a concept to keep it correct and easy to understand. T1 took the incorrect analogy of "nuclear fission and fusion" (Fig. 3) to illustrate LLMs' difficulty in generating correct physical analogies, noting that forcing analogies for all features leads to factual and semantic errors. He explained that physical concepts often involve multiple features, some of which, like "chain reactions", can be analogized (e.g., "dominoes"), while others, such as "mass-energy conversion", are too abstract to find counterparts due to their basis in mathematical models. For biological analogies, T2 recommended focusing on negative feedback in "thyroid hormone regulation" with an analogy like "adjusting the temperature with an air conditioner remote control," while students should memorize other details. S2 recalled an analogy about specific details, in which the teacher compared a "channel protein" with a "fire escape." Therefore, for complex concepts with multiple knowledge points, selecting only a specific aspect for the analogy is sufficient.

Not necessary to generate perfect analogies. Teachers were lenient towards the generated analogies from Study I and managed to extract effective parts from them. T2 appreciated the analogy comparing "nerve impulses" to the "efficient operation of stations in an express delivery system," though some parts were redundant. Additionally, T1 shared his experience using ChatGPT for lesson plans, finding it repetitive and sometimes vague but useful for providing new ideas.

Evaluating LLM-generated analogies in class by teachers. Teachers had various approaches to evaluating the effectiveness of analogies in class. T1 asked questions like "Have you seen something similar before?" or observed the students' expressions, while T2 had the students answer concepts-related questions during class. Additionally, two teachers expressed cautious optimism about using LLM-generated analogies with their interventions. T1 noted that frequently used analogies for physics were concept-based, while AI-generated ones felt more relatable to everyday life, which makes him uncertain about their actual effects. Besides, both T1 and T2 anticipated better classroom feedback but were unsure of the effects on students' performance on homework and exams. This led to a consensus on conducting a comparative experiment.

5.2 Classroom Experiments

In this subsection, we describe the participants (Sec. 5.2.1), data preparation process (Sec. 5.2.2), procedure (Sec. 5.1.3), and results analysis (Sec. 5.2.4) for our classroom experiments.

Table 3: A summary of interviewing physics and biology teachers.

Topic	Physics Teacher (T1)	Biology Teacher (T2)
Analogy Usage Exploration		
Analogies Frequency	Sometimes.	Frequent.
Analogies Feature	Mostly between learned concepts.	Mostly between biology and daily life.
Source of Analogies	Mostly Prepared analogies between concepts. A few improvised analogies with everyday lives.	Mostly prepared analogies. Nearly no improvised analogies.
Good Analogy Criterion	Easy to understand and free of scientific errors.	Easy to understand and related to everyday life.
Agreement with Initial Principles in Study I	Partial agreement: Analogies between similar physical concepts.	Total Agreement.
Analogy Explanation	Verbal explanation + imagery + teaching aids	Verbal explanation + imagery + teaching aids
Analogies Usage Scenario	Often used to introduce concepts. Sometimes throughout teaching.	Often used when detailing knowledge points.
Agreement with the Six-step Model of Practice [62]	Acknowledges most, except for pointing out differences when introducing concepts.	Total agreement.
Student Participation in Constructing Analogies	Rare. Sometimes, students offer their ideas, which might be used in the next class.	Rare. Sometimes, students prepare analogies for student-led discussions.
Students Understanding Examination	Question students with "Have you seen something similar before?", or observe students' expressions	Students complete a few exercises during class, or question students about concept differentiation.
AI Usage Exploration		
Awareness and Experience with AI	Has used ChatGPT for writing papers, lesson plans, and creating images; knows about Sora.	Has used ChatGPT for tenders and personal use.
Pros and Cons of AI	Pros: helps write unexpected things. Cons: Needs specific questions; AI usually doesn't follow the instructions.	Pros: Provides broad ideas. No clear cons due to limited experience.
Can AI Replace Teachers?	Teachers know students' learning situations, AI does not; AI-generated content needs adjustment.	AI cannot replace but complement teachers.
Expectations Sharing on LLM-generated Analogies		
Positive Comments on Analogies in Study I	1. The analogies are all vivid and some of them are interesting	1. Some analogies are similar to those used in class 2. Identify analogies to try in class for concepts not usually taught with analogies.
Negative Comments on Analogies in Study I	1. Analogies don't clarify abstract concepts. 2. Analogies can complicate simple concepts. 3. For concepts that are tested simply, memorization is enough. 4. Pictures could make some concepts clear without analogies.	1. Analogies shouldn't reflect all but the main concepts; the rest relies on memory. 2. Pictures and animations can visualize familiar organisms without analogies. 3. Although rare, related concepts sometimes are used as analogies.
Overall Expectations	Vivid analogies between physical concepts.	Analogies from daily life for teaching focus; Interesting analogies to stimulate learning interest.

5.2.1 Participants. In this one-week field study, participants included two teachers (T1 and T2) from the pre-class interviews and two first-year high school classes (C1 and C2) they were teaching. Each class had 25 students, 12 of whom were girls, and the distribution of their entrance exam scores was very similar.

5.2.2 Data Preparation. Teachers informed us about concepts that might require analogies in the following week of teaching. The concepts taught by the physics teacher (T1) include average velocity and instantaneous velocity, acceleration, and infinitesimal method. The concepts taught by the biology teacher (T2) include the various functions of proteins, the adaptation of function and structure, dehydration condensation, the formation of tertiary and quaternary structures, and protein denaturation. Based on pre-class interviews, we identify four effective strategies to generate analogies from LLMs for classroom practice.

- **Strategy 1: Analogy for Physical Concept.** For physical concepts, analogies often draw on learned physical concepts.
- **Strategy 2: Analogy for Biological Concept.** For biological concepts, analogies often involve everyday objects. For example, one might use the analogy of fire escape to help understand channel protein.

- **Strategy 3: Vivid Analogy Generation.** Analogies should be vivid and engaging to capture students' attention.
- **Strategy 4: Fine-grained Analogy Generation.** Sometimes, it is sufficient to generate analogies for just one aspect of a concept to provide a detailed explanation of that particular aspect.

Based on the strategies outlined, we can modify the prompt in Tab. 1 to suit the specific aspect of a concept and the requirements of teachers. Specifically, we incorporated Strategies 1, 2, and 3 into the *Principles*. Strategy 4 was added into the *Input Resource* to prevent the model from forgetting.

Following discussions with T1, we generated analogies for average and instantaneous velocity by implementing either strategy 1 or 3. Additionally, we generated detailed analogies for the infinitesimal method and acceleration using strategy 4. In biology, we produced analogies for proteins by applying either strategy 2 or strategy 3. Using strategy 4, we developed detailed analogies for the immune effects of proteins and the formation of tertiary and quaternary structures. However, two generated analogies, "driving speed" and "reading speed", were marked as non-analogies and excluded. Besides, the analogies generated with Strategy 1 for physical concepts were not related to other concepts, but we included them as they are vivid analogies. We generated four physical analogies to T1 and nine biological ones to T2.

5.2.3 Procedure. In this one-week teaching, T1 and T2 used LLM-generated analogies for C1 and kept the original teaching mode for C2, with each class having 3 lessons for each subject, totaling 12 lessons. One author attended one C1 lesson taught by T1 and one by T2, observing how teachers used analogies and student reactions without disrupting teaching. For the remaining lessons within the week, teachers completed our provided record forms after each lesson. The record forms asked for details on which analogies they chose while preparing for C1, any modifications made to these analogies, and reasons for not selecting others. Additionally, the forms inquired about how teachers assessed student feedback during or after class, any differences in feedback between C1 and C2, and whether the feedback met their expectations. After one week, we conducted one-on-one interviews with T1 and T2, each lasting 20 minutes, to clarify any unclear details in the records, and discuss their experiences with LLM-generated analogies, students' performance, and future expectations. Both teachers received a $60 gift card each for their dedicated participation over the week.

5.2.4 Results Analysis. We report the following qualitative findings based on the record forms and interviews.

Teachers selected and modified LLM-generated analogies to avoid redundancy, confusion, or misleading students and make them closer to students' daily lives. T1 selected two of four analogies and modified one, while T2 chose four of nine analogies and modified two. In the interview, T2 noted that while the analogies for all four functions of proteins had merits, only two were selected to avoid verbosity in the class. Besides, to avoid concept confusion, T2 chose the analogy of "transport function" as a "conveyor belt" and discarded the analogy of "catalysis function" as a "high-speed elevator," due to the transport function of the elevator. We observed two types of modifications made by teachers to analogies. One type involved modifying details, such as T2 changing the security guard's action from "eliminating" to "capturing" to align with the real-world context (Fig. 6B). Another type involved changing analogy objects, like T1 replacing "jigsaw puzzle" with "pixels on a display screen" to illustrate the infinitesimal method (Fig. 6C). T1 explained that display screens are more familiar to students than jigsaw puzzles.

LLM-generated analogies inspire teachers with new analogies and new teaching methods. In the interview, both teachers recognized the novelty of some LLM-generated analogies for concepts, and they had not considered using analogies for those concepts before. For example, T1 used "video and snapshot" to analogize "average velocity and instantaneous velocity" (Fig. 6A), while T2 used "wool folding and weaving" to analogize the "tertiary and quaternary structures of protein." In addition, T2 developed new analogies inspired by LLM-generated analogies. While teaching the "dehydration condensation reaction," T2 explained with a new analogy as "breaking down the wall between classrooms" (Fig. 6D). In the interview, T2 said, "*I am not satisfied with the generated one, as comparing the dehydration condensation reaction to mixing building materials doesn't capture the essence. However, the buldings environment inspired me to create a new analogy.*" Besides, T1 said that participating in this study had changed his teaching style. T1 used analogies based on everyday life after explaining the concept, which was inconsistent with his pre-class interview response.

LLM-generated analogies boost students' classroom and homework performance and encourage teaching with analogy. Both teachers believe that C1 outperforms C2 in both classroom participation and homework. T1 praised analogies for helping students focus on the class: "*I can see from the students' eyes that C1 is genuinely paying closer attention, with more students nodding sincerely, rather than just pretending.*" T1 also reported that C1 outscored C2 by nearly 20% on a 10-question homework. He attributed this to C2's confusion between average and instantaneous velocity, causing errors on the two hardest questions." In the interview, T1 said, "*I plan to use the analogies in C1 when reviewing the assignments in C2 to explain the concepts again.*" As for biological concepts, T2 said, "*When I explained protein structure using a video, C2 students understood initially but got confused about the tertiary structure, whereas the wool stacking analogy helped C1 students understand the video.*" T2 showed us a fill-in-the-blank question from homework that asked students to summarize protein function. Most C1 students summarized correctly, while many C2 students simply copied words from the textbook. However, T2 noted that there was no clear difference between the two classes in understanding straightforward concepts like "Protein denaturation". T1 also noted that the physics analogies are still not between concepts and may offer limited help with highly abstract concepts, while he added "*But I'll try more teaching with analogy since the difference between the two classes is clear.*"

Overall, promising feedback from teachers and classroom practice led us to consider designing a practical system to support the preparation of teaching analogies.

6 System

In this section, we transformed key study findings into an LLM-assisted system for teachers and conducted a system evaluation, highlighting its contribution to teaching by analogy in education.

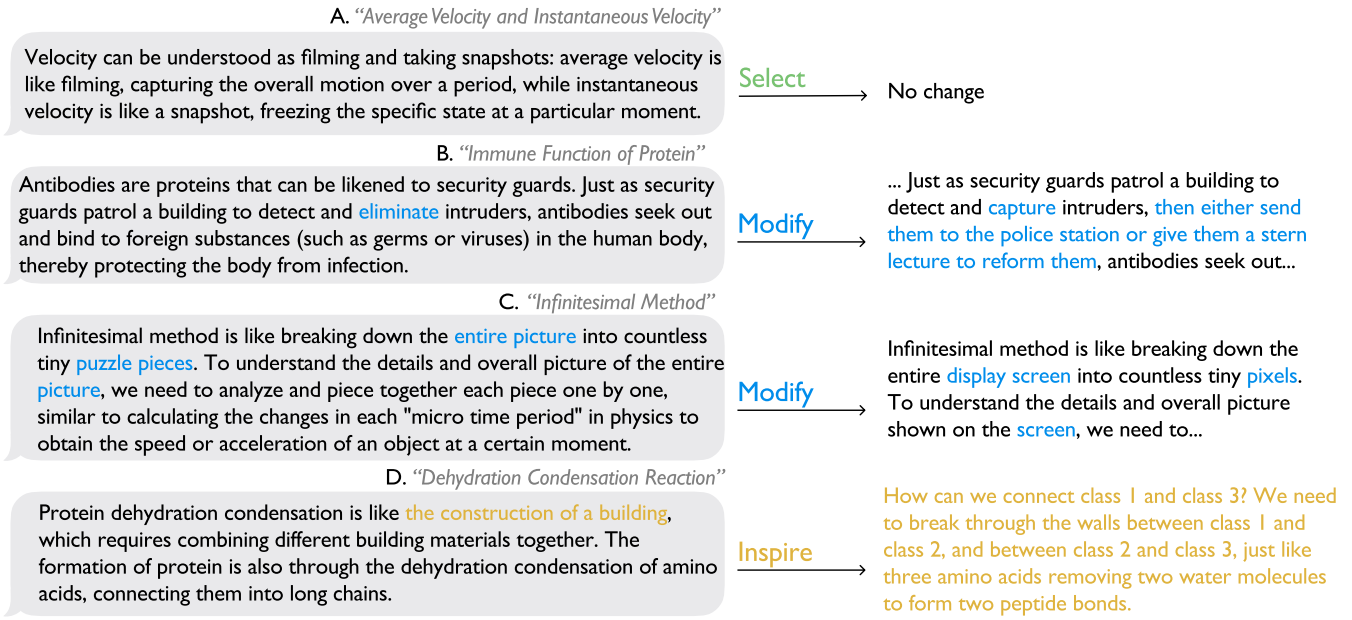

Figure 6: Teachers' behaviors on LLM-generated analogies. They may either directly select (A) and use them in class, modify their details (B) or analogy objects (C) to varying extents, or even create entirely new analogies inspired by them (D).

6.1 System Design

This subsection details the interview for deriving design requirements and system workflow as below.

6.1.1 Interview. Given that Study II showed teachers prepared analogies during lesson planning, we first determined the system's necessity and functionality through 20-minute one-on-one online interviews with T1 and T2 via Tencent Meeting. The interview mainly consists of two questions. 1) Necessity: Is providing an LLM-assisted analogy generation system necessary for teachers to prepare lessons? 2) Functionality: What functions do they expect?

Both teachers affirmed the first question, expressing a desire to operate the system themselves to gain hands-on experience and long-term support from LLMs. Regarding the second question, both teachers expressed a preference for the analogy generation mode in Study II. They suggested that after specifying the required concepts, the system should generate accurate analogies tailored to their needs, allowing for refinement and management. They also noted that in Study II, analogies function as plug-and-play modules to replace or enhance original explanations of scientific concepts, so the generation process need not account for other lesson plan content at this stage.

6.1.2 Design Requirements. After confirming the necessity and functionality, we identified three design requirements as below and confirmed them with T1 and T2.

R1: It should incorporate principles and strategies identified in previous studies to help generate accurate analogies tailored to teacher needs. The general principles identified in Study I (Tab. 1) should be integrated into the prompt by default to enhance the accuracy. The system should allow teachers to select useful prompting strategies identified in Study II and incorporate them into generation following the practice of data preparation of Study II (Sec. 5.2.2) to better tailor analogies to their needs.

R2: It should enable teachers to input their expectations or automatically generate personalized principles for creating analogies. As identified in Study II interviews (Sec. 5.1.3), the workflow should allow teachers to input concepts they believe require analogies to be generated. Besides, to ensure personalized needs, it should also accept user-inputted new principles tailored to their needs and even automatically generate tailored principles from user comments on generated analogies for the prompt (Tab. 1).

R3: It should allow users to make manual changes and feedback to manage their analogies. In Study II, teachers showed a strong willingness and ability to refine and manage generated analogies (Sec. 5.2.4), highlighting the need for a system that allows direct editing. As prompt evolution in Study I showed limited improvement (Tab. 2), and teachers found conversational refinement challenging in pre-class interviews in Study II (Tab. 3), conversational modifications by LLMs are unnecessary. The system should also enable teachers to manage analogies by classifying them into four categories (useful, inspiring, refinable, and useless) and storing all except the useless ones.

6.1.3 System Workflow. Based on the design requirements, we built an LLM-assisted system (Fig. 7) for teachers to create and refine analogies for teaching. Teachers begin by registering an account and follow a workflow as below.

Configure strategies and principles for analogy generation. After selecting a teaching subject during registration, the system provides prompting strategies in Configuration Panel (Fig. 7A) based on the chosen subject. Teachers can click on these strategies for generating analogies (**R1**). The principle list starts blank, allowing teachers to manually add or select principles as needed (**R2**).

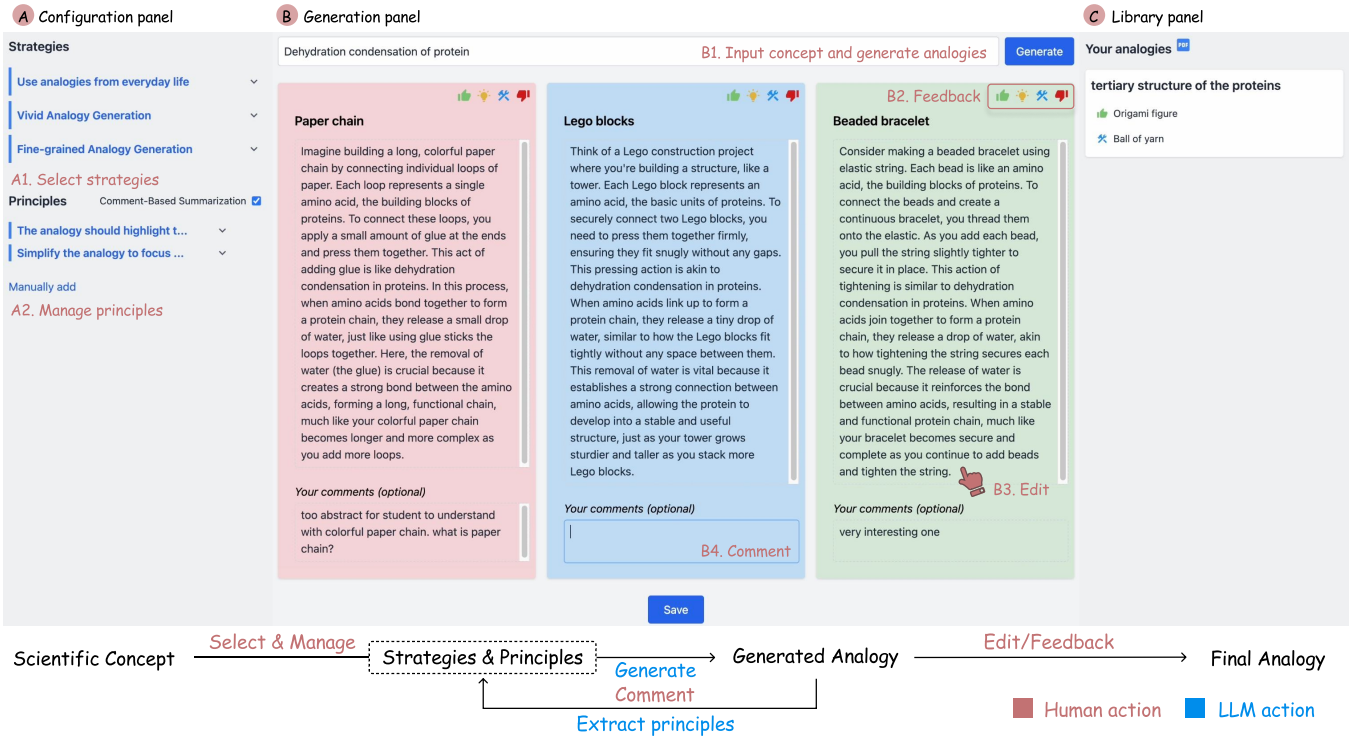

Figure 7: Our system interface (top) and workflow (bottom). In each round, teachers use Configuration Panel (A) to select strategies (A1) and manage principles (A2) for generation. After entering scientific concepts (B1), they provide feedback (B2), edit (B3), and comment (B4) to each generated analogy in Generation Panel (B). Clicking the "Save" button makes the system generate new principles by LLM in Configuration Panel and store approved analogies in Library Panel (C), where teachers may export saved analogies.

When the user hovers over a principle, they could edit or delete it freely.

Generate analogies and provide feedback. Teachers then input a scientific concept in Generation Panel (Fig. 7B) and click the "Generate" button (**R2**). The system will provide three cards with generated analogies. Each card has four feedback options in the top-right corner: useful, inspiring, refinable, and useless. Teachers need to select one feedback option for each analogy card and may edit analogies as needed and provide text comments (**R3**).

Save analogies, optionally generate new principles, and restart. When saving, modified analogies (excluding those marked as useless) are stored in Library Panel (Fig. 7C), along with the corresponding concept (**R3**). If teachers enable the "Comment-based summarization" feature in Configuration Panel (Fig. 7A), the system will summarize new principles from comments on the three analogies by LLMs and add them to the principles list (**R2**). Teachers can edit or delete any generated principles or disable the automatic summarization feature at any time. After saving, the system clears the concepts and analogies in Generation Panel (Fig. 7B), allowing users to restart. Teachers can reconfigure the strategy list, and add, delete, select, or modify the principle list for the new round.

Besides, teachers click right buttons of text for strategies (Fig. 7A1), principles (Fig. 7A2), and approved analogies (Fig. 7C) to view or collapse text, reducing clutter. All approved analogies stored in Library Panel can be exported to PDF for teachers' usage.

We implemented the system as a Vue-based web application with a Python Flask backend. More implementation details, such as prompts for automatically generated principles, are provided in the supplementary material.

6.2 System Evaluation

This subsection details the participants, procedure, and findings of the system evaluation.

6.2.1 Participants. We invited 6 high school physics and biology teachers from 5 schools, including T1 and T2, along with 4 new teachers from different schools to increase diversity. The physics teachers had 7 years (T1), 3 years (T3), and 4 years (T5) of experience, while the biology teachers had 4 years (T2), 2 years (T4), and 25 years (T6) of experience. All participants received a $20 gift card as compensation.

6.2.2 Procedure. Our study consists of a tutorial, free exploration, one-week usage, and an interview.

Tutorial. We conducted one-on-one online interviews with each teacher other than T1 and T2 via Tencent Meeting, lasting 20 minutes. We briefly revisited Steps 1 and 3 from Study II (Sec. 5.1) to understand their experience with analogical teaching and AI, and to ensure they were familiar with the study background.

Free exploration. Following the tutorial, we continued the interview with each teacher to demonstrated the basic functionality

of the system according to the workflow, after which the teachers were encouraged to explore the system freely. During this exploration phase, we prompted them to think aloud and we answered any questions they had to make sure they understand how to use the system. This process lasted about 20 minutes.

One-week usage. Before using the system, we informed the teachers that it would collect data on the analogies and principles they generated, as well as their feedback, manual edits, comments, and other interaction data for research purposes. All of them provided informed consent. We then asked the teachers to use the system to generate analogies to support their lesson preparation for one week. The frequency of system use and the content of lesson preparation were determined by the teachers themselves, although we encouraged them to use the system frequently.

Interview. After one-week usage, we conducted one-on-one semi-structured online interviews with each teacher via Tencent Meeting, lasting nearly 40 minutes each. During the interviews, we asked each teacher the following questions: course coverage (Q1), system usability (Q2), satisfaction with the generated analogies (Q3), satisfaction with the generated principles (Q4), satisfaction with the edited analogies (Q5), satisfaction with the edited principles (Q6), and the significance of the system for their future lesson preparation and teaching (Q7).

6.2.3 Findings. Based on user interaction data and interview results of Q1-Q7, we summarize the following findings.

The system's usability and the usefulness of generated analogies were well-received by teachers (Q1-Q3). Teachers reported that they used the system to generate analogies for lessons spanning from one month to half a semester, aiding both reflection on past lessons and preparation for upcoming ones. All teachers agreed that the system was easy to use, even if they had nearly no experience using AI (T3, T5, T6). They generated 15 to 42 analogies (15, 15, 21, 24, 24, 42), with 40% to 80% marked as non-useless ("useful", "refinable", "inspiring") across users (40%, 58.3%, 61.9%, 64.3%, 66.7%, 80%). They all agreed that although many of the generated analogies had various issues, overall, they helped expand their lesson planning ideas and inspired greater use of analogies in teaching. For example, the system analogizes "phase difference" to "some students doing radio gymnastics faster or slower than classmates." T3 said, "*I hadn't thought of that, but it's great, and since my class is right after gymnastics session, I'll definitely use it.*" Besides, physics teachers noted that they understood the system's inability to provide suitable analogies for some complex concepts, as such analogies might not exist anyway.

Teachers tried to improve analogy quality by incorporating generated principles or directly regenerating analogies (Q4). Five teachers (excluding T4) continuously enabled the automatic principle generation feature and actively input their comments, producing 10 to 30 principles (10, 13, 17, 18, 30) and incorporating at least half into analogy generation. They all praised the quality of the principles, with T1 saying, "*The system infers general principles from my vague intents on comments on specific analogies.*" For usability of this feature, T3 noted, "*It doesn't matter if too many principles are generated during the usage. I just delete the redundant ones—it's more convenient than summarizing myself.*" For its effectiveness in analogy generation, T1 and T5 felt the principles improved analogy quality, while T3 and T6 were unsure but still incorporated them because "*adding more correct terms can't hurt.*" On the contrary, T4 preferred regenerating analogies directly without commenting on analogies and generating principles, saying, "*If the analogies aren't good, I just regenerate them a few more times as I know the randomness of AI.*" Only T4 manually added principles at the beginning of usage, while all teachers found it inconvenient and difficult as noted by T3 and T1.

The generated analogies and principles benefit teachers more than just concept explanations (Q7). 1) The principles help shape teaching expertise. T1 said, "*I can learn teaching techniques from the generated principles, and after using the system for a while, I could even write a teaching paper. It's like having a discussion about teaching with another experienced teacher. While the principles may not immediately impact teaching, the long-term accumulation is valuable.*" 2) The analogies supplement the teacher's knowledge base. T6 noted that the generated analogies broaden a teacher's perspective, and saving more analogies gradually builds their teaching knowledge system, which help teachers adapt to different teaching situations. 3) The analogies also inspire quiz and test creation. T2 said, "*Even if some analogies aren't ideal, I save them because having students identify errors helps their learning.*" T3 stated, "*Some of the generated 'analogies' is example-based explanation, but I save it because these real-life examples can be used to set questions.*"

Teachers typically organize analogies externally rather than refining them within the system and there is potential over-reliance (Q5-Q6). Despite marking large proportion of generated analogies as "refinable" (Mean = 21.4%) and "inspiring" (Mean = 13.8%), only T1 and T4 edited 1 analogy in the system, respectively. T6 explained, "*Manually modifying so much text is too burdensome for older teachers.*" However, all teachers reported improving analogies to fit their needs through external modifications. T5 noted that his lesson preparation habit is to record keywords in electronic notes, so the analogies in the system only serve as explanation and inspiration, which he then reorganizes in his notes. Similarly, T1 and T6 preferred recording analogies in paper notebooks. This separation between analogy generation and actual lesson preparation may make it difficult to supervise teachers' behaviors in teaching and potentially lead to teachers' over-reliance on generated analogies. This issue may be more pronounced for users like T4, who prefer to regenerate analogies directly without providing any comment, compared to teachers who actively engage by entering comments in the system.

7 Discussion

This section discusses consideration, opportunities, and future research directions for LLM-assisted analogical education based on our study results and designed system.

7.1 Subject Differences in LLM-Generated Analogy Effectiveness

Our study shows that LLMs generally produce correct and satisfying analogies for biological concepts but generate incorrect or correct yet unprofessional ones for physics.

Several factors appear to contribute to these shortcomings. First, physical concepts are highly abstract (e.g., "mass-energy conversion"), with complex and formula-driven features, making it difficult to find real-life analogies or other concepts that perfectly align with them. In contrast, biological concepts are more concrete and observable, often tied to specific structures and functions (e.g., "mitochondria as the powerhouse of the cell"). As a result, LLMs would produce forced analogies or oversimplifications for physics while generating satisfying ones for biology. Second, restricting the physics analogy to a single aspect, as in the strategies used in Study II (Sec. 5.2.2), can yield correct and engaging analogies. However, for highly abstract concepts, these analogies may still be superficial and offer limited support, as noted by physics teachers in Sec. 5.2.4. Third, teaching materials in physics contain more formulas and fewer analogies compared to biology. As a result, LLMs learn fewer physics analogies and generate less effective analogies.

These findings suggest that using LLMs for educational analogy generation is tied to subject characteristics, and we can infer that it may be particularly challenging for subjects lacking clear real-world counterparts (e.g., mathematics). In contrast, they might work better for subjects with more directly observable phenomena (e.g., high school chemistry, biology), which should be confirmed by future studies. Nonetheless, analogies help students engage with abstract subjects like physics and math by inspiring interest and sustaining attention. More studies are needed to verify the effectiveness and needs of LLM-generated analogies across a broader range of subjects, in conjunction with the review and refinement of teachers before their use.

7.2 Generating High-Quality Analogies

The automatically generated analogies have limitations in scientific accuracy and educational effectiveness. In our practical system, LLMs summarized principles from human feedback and incorporated them into the next round of analogy generation. Several teachers in the system evaluation found this approach effective for improving analogy quality. Building on this, we can incorporate Reinforcement Learning with Human Feedback (RLHF). Using teachers' feedback and preferences, we can create a reward model that continually refines the analogies. To further reduce teachers' workload, future work should explore automatic methods for generating higher-quality analogies. First, we could explore multi-agent collaboration methods to further mitigate hallucinations [65], including factuality errors and consistency errors, as outlined in Study I. Besides, instead of waiting for more advanced general-purpose LLMs to be released [55], we can fine-tune existing models with teacher-adjusted analogies. Additionally, our system can be transformed into a labeling tool to collect high-quality educational analogy datasets, consisting of revised analogies or new analogies proposed by teachers. Another bottleneck for generating high-quality analogies is the LLMs' limited domain-specific knowledge. In Study II, the physics teacher attributed the current analogies' interesting yet unprofessional nature to the LLMs' limited understanding of abstract physics concepts. To address this, we should fine-tune LLMs for specific subjects, improving their understanding and enhancing analogy quality. Finally, to improve control over the complexity and ethical considerations and make analogies suited for the intended educational level and scenario, the future analogy generation pipeline should consider factors like students' educational background, cultural context, and prior knowledge.

7.3 LLMs for Teaching by Analogy

Our Study II and system evaluation show that LLM-generated analogies are valuable to teachers in three progressive aspects: short-term lesson preparation, teaching strategy development, and professional growth. For short-term lesson preparation, teachers are able and willing to select and modify LLM-generated analogies or inspire new ones to suit specific concepts and lessons (Sec. 6.2.3), or use generated content to help set quizzes (Sec. 5.2.4). For teaching strategy development, continuous use of LLM-generated analogies leads to positive feedback from the classroom and students' homework and iteratively encourages teaching by analogy (Sec. 5.2.4). Regarding professional growth (Sec. 6.2.3), providing feedback on LLM-generated analogies helps teachers actively reflect on teaching points and build their knowledge base, while LLM-generated principles based on their feedback also serve as valuable reminders for teachers, supporting their ongoing professional development and enhancing teaching expertise. Given these benefits, future work should explore the varying needs and develop practical systems to benefit teachers with different experience levels and subjects. In addition, long-term evaluation of such practical systems and teachers is needed to fully understand the actual benefits.

Besides the benefits, over-reliance on LLM-generated analogies warrants attention. In Study II, teachers emphasized during preclass interviews and demonstrated in classroom teaching that they could avoid over-relying on such content. However, in the system evaluation, most teachers did not revise analogies within the system but recorded changes elsewhere, following their own lesson preparation habits. This suggests that monitoring teachers' interactions with analogies only through system logs might be insufficient, potentially allowing unnoticed over-reliance to develop. To address this, the system could use pop-up reminders to alert users against over-reliance and encourage them to edit or provide comments when no manual interaction is detected for an extended period. Nevertheless, supervision from schools and higher authorities is essential. Additionally, regular updates from system developers to educators [40, 69] are crucial for maintaining an accurate understanding of model capabilities and ensuring effective use.

7.4 Integrating Analogies into LLM-Assisted Education Platforms

Integrating analogy generation into LLM-assisted education platforms might benefit teachers and students.

For teachers, the interplay between analogy generation and LLM-assisted teaching material preparation is mutually reinforcing. First, analogies help develop teaching materials by providing relatable explanations. For example, LLM-assisted platforms already help novice teachers generate lesson plans [17]. Integrating analogy generation into these platforms can support analogy-based explanations at different teaching stages. Additionally, participants in our system evaluation demonstrated the potential of using analogies in quiz settings, highlighting its role in automated problem generation. Second, existing LLM-assisted teaching preparation

platforms enhance context-aware analogy generation and verification. These platforms usually consider students' knowledge levels and course context [45], which could be integrated into analogy generation pipelines to help generate analogies suited for practical scenarios. Moreover, existing problem generation platforms could be enhanced to generate quizzes that verify students' understanding of analogy, reinforcing the effectiveness of teaching by analogy.

However, the results of Study I indicate that LLM-generated analogies without human intervention are unreliable for students, making it premature to directly integrate analogy generation into self-learning systems. In contrast, teacher-adjusted analogies in Study II ensured correctness and reliably impacted students' classroom feedback and homework performance. Therefore, future self-learning systems should only consider pre-set teacher-reviewed analogies for key knowledge points to aid student understanding. However, even correct biological analogies from Study I led to negative effects, with students over-relying on them with incorrect learning strategies or becoming subjectively overconfident. This suggests that the self-learning system should flexibly structure the learning process and, after presenting analogies, guide students back to the textbook with more detailed follow-up questions. Besides, timely pre-set and teacher-reviewed exercises with feedback and explanation that reveal the limitations of analogies help students reflect on their learning and monitor their learning approaches. Overall, incorporating analogies into self-learning systems requires careful attention from teachers and system developers to mitigate potential negative effects.

7.5 Evaluating Broader Analogies in Education

Analogies serve multiple purposes beyond students' understanding and teachers' teaching, which adds challenges to their evaluation. In mathematical problems and similar domains, specific procedures involving numeracy and variables often require a different type of analogy, known as procedural analogy [61]. Such procedural analogies were also mentioned during our interview with the physics teacher in Study II. Due to their rarity and complexity, these analogies, even those crafted by humans, have not been thoroughly evaluated. LLMs can lower the barrier to creating such analogies due to advanced reasoning abilities and broad subject knowledge, such as linking the formula for a "spring oscillator" with that of a "pendulum." Our study design can be extended to such analogies by involving calculation questions involving formulas in controlled in-class tests. Additionally, analogies are utilized for socialization, helping to educate children on becoming better students and enacting behavioral changes [61]. The evaluation of such analogies involves contexts beyond the classroom, which brings challenges to study design and needs to be explored in the future.

7.6 Limitation

Although we have gained lots of evidence and knowledge, our work is limited by student and teacher participation and analogy representation.

7.6.1 Limitations in Student and Teacher Participation. Due to practical limitations, the teachers and students in our first two studies were from one high school, and the sample size was limited. To explore real-world practicality, we expanded the participant pool by inviting more teachers from different schools with varying teaching experiences to evaluate the practical system. Additionally, our study is limited to physics and biology due to practical constraints, excluding other subjects like chemistry. Besides, our study with high school students may not be generalizable to younger students who might not possess developed analogical reasoning abilities or have limited world knowledge [76]. Moreover, our study's demographic generalizability is limited, as all participants are Chinese, while prior research [63] suggests that U.S. teachers provide cognitive support for analogies less frequently than teachers from Hong Kong or Japan in math instruction. In future work, we intend to expand the sample size, range, and diversity of subjects and explore diverse education levels for comprehensive evaluation.

7.6.2 Limitations in Representation of Analogy. The practical use of analogies in teaching extends beyond the free-form analogies we generated, incorporating visual aids and dynamic technologies to enhance understanding and interaction [62, 63]. Our interviews in Study II and system evaluation also revealed that teachers have the desire to use images, videos, and physical aids to convey analogies. Moving forward, we plan to enrich LLM-generated analogies with rich text, structured representations, and generated visuals to benefit teachers in practical systems.

8 Conclusion

In this work, we first conducted in-class tests and classroom experiments guided by pre-class interviews to evaluate the effectiveness of LLM-generated analogies in two educational scenarios. Our in-class tests suggest that LLM-generated analogies could be beneficial for students' understanding, especially on biology concepts, but are unsuitable for self-learning systems without teacher intervention due to students' over-reliance and overconfidence. Classroom experiments reveal that teachers effectively create or refine analogies to meet their needs and are encouraged to teach with analogy by positive student feedback in class and homework. Building on these findings, we developed a practical system for teachers preparing analogies. Teachers in the system evaluation recognize its real-world effectiveness in lesson preparation about concept explanation and quiz design, teaching methods, and professional growth, despite potential over-reliance issues. We hope future tool designers could consider these factors to ensure that LLM-generated analogies have a successful impact on teaching and learning.

Acknowledgments

The authors want to thank the reviewers for their suggestions. The authors also thank SHANGHAI CUNZHI HIGH SCHOOL for its generous support and all participating teachers and students, especially Siwei Ye and Zihan Luo for their deep involvement, and Yuhui Wang for strong support. This work is supported by Natural Science Foundation of China (NSFC No.62472099, No.62202105, and No.92270121).

References

[1] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk)

(CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300760

[2] Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. In *Proceedings of the 15th International Conference on Natural Language Generation*. Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting, 298–312. https://aclanthology.org/2022.inlg-main.25

[3] Bhavya Bhavya, Yang Zhou, Shradha Sehgal, Suma Bhat, and ChengXiang Zhai. 2024. Analego: Let's build analogies together!. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*.

[4] Adrian Boteanu and Sonia Chernova. 2015. Solving and Explaining Analogy Questions Using Semantic Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (Feb. 2015). https://doi.org/10.1609/aaai.v29i1.9400

[5] Ann L. Brown, Mary Jo Kane, and Carolyn Long. 1989. Analogical transfer in young children: Analogies as tools for communication and exposition. *Applied Cognitive Psychology* 3, 4 (1989), 275–293. https://doi.org/10.1002/acp.2350030402 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350030402.

[6] David Casarett, Amy Pickard, Jessica M. Fishman, Stewart C. Alexander, Robert M. Arnold, Kathryn I. Pollak, and James A. Tulsky. 2010. Can Metaphors and Analogies Improve Communication with Seriously Ill Patients? *Journal of Palliative Medicine* 13, 3 (March 2010), 255–260. https://doi.org/10.1089/jpm.2009.0221

[7] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (nov 2018), 21 pages. https://doi.org/10.1145/3274300

[8] Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 3941–3955. https://doi.org/10.18653/v1/2022.findings-acl.311

[9] Liuqing Chen, Zhaojun Jiang, Duowei Xia, Zebin Cai, Lingyun Sun, Peter Childs, and Haoyu Zuo. 2024. BIDTrainer: An LLMs-driven Education Tool for Enhancing the Understanding and Reasoning in Bio-inspired Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 676, 20 pages. https://doi.org/10.1145/3613904.3642887

[10] Qing Chen, Wei Shuai, Jiyao Zhang, Zhida Sun, and Nan Cao. 2024. Beyond Numbers: Creating Analogies to Enhance Data Comprehension and Communication with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3613904.3642480

[11] Zixin Chen, Jiachen Wang, Meng Xia, Kento Shigyo, Dingdong Liu, Rong Zhang, and Huamin Qu. 2024. StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions. *arXiv preprint arXiv:2407.12423* (2024).

[12] Todd Davies. 1985. Analogy. *CSLI Informal Notes Series IN-CSLI-85-4, Center for the Study of Language and Information, Stanford* (1985).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[15] Zijian Ding, Arvind Srinivasan, Stephen Macneil, and Joel Chan. 2023. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. In *Proceedings of the 15th Conference on Creativity and Cognition (C&C '23)*. Association for Computing Machinery, New York, NY, USA, 489–505. https://doi.org/10.1145/3591196.3593516

[16] Reinders Duit et al. 1991. On the role of analogies and metaphors in learning science. *Science education* 75, 6 (1991), 649–672.

[17] Haoxiang Fan, Guanzheng Chen, Xingbo Wang, and Zhenhui Peng. 2024. LessonPlanner: Assisting Novice Teachers to Prepare Pedagogy-Driven Lesson Plans with Large Language Models. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 146, 20 pages. https://doi.org/10.1145/3654777.3676390

[18] Mirta Galesic and Rocio Garcia-Retamero. 2013. Using Analogies to Communicate Information about Health Risks. *Applied Cognitive Psychology* 27, 1 (2013), 33–42. https://doi.org/10.1002/acp.2866 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2866.

[19] Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chiokit Ieong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. 2024. Fine-Tuned Large Language Model for Visualization System: A Study on Self-Regulated Learning in Education. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–11. https://doi.org/10.1109/TVCG.2024.3456145

[20] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

[21] Dedre Gentner and Kenneth D Forbus. 2011. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science* 2, 3 (2011), 266–276. Publisher: Wiley Online Library.

[22] Dedre Gentner and Christian Hoyos. 2017. Analogy and abstraction. *Topics in cognitive science* 9, 3 (2017), 672–693.

[23] Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist* 52, 1 (1997), 45. Publisher: American Psychological Association.

[24] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. https://doi.org/10.1145/3613904.3642139

[25] Mary L. Gick and Keith J. Holyoak. 1980. Analogical problem solving. *Cognitive Psychology* 12, 3 (July 1980), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

[26] Mary L. Gick and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (Jan. 1983), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

[27] Karni Gilon, Joel Chan, Felicia Y. Ng, Hila Liifshitz-Assaf, Aniket Kittur, and Dafna Shahaf. 2018. Analogy Mining for Specific Design Needs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173695

[28] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 8–15. https://doi.org/10.18653/v1/N16-2002

[29] Ashok K Goel. 1997. Design, analogy, and creativity. *IEEE expert* 12, 3 (1997), 62–70. Publisher: IEEE.

[30] Maureen E. Gray and Keith J. Holyoak. 2021. Teaching by Analogy: From Theory to Practice. *Mind, Brain, and Education* 15, 3 (2021), 250–263. https://doi.org/10.1111/mbe.12288 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mbe.12288.

[31] Mary B Hesse. 1959. On defining analogy. In *Proceedings of the Aristotelian Society*, Vol. 60. JSTOR, 79–100.

[32] Rostyslav Hnatyshyn, Jiayi Hong, Ross Maciejewski, Christopher Norby, and Carlo C. Maley. 2024. Capturing Cancer as Music: Cancer Mechanisms Expressed through Musification. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 727, 11 pages. https://doi.org/10.1145/3613904.3642153

[33] Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

[34] Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. In-Context Analogical Reasoning with Pre-Trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 1953–1969. https://doi.org/10.18653/v1/2023.acl-long.109

[35] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving Comprehension of Measurements Using Concrete Re-expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173608

[36] Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11518–11537. https://doi.org/10.18653/v1/2023.emnlp-main.706

[37] Hyeonsu B Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. 2024. BioSpark: An End-to-End Generative System for Biological-Analogical Inspirations and Ideation. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613905.3651035

[38] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Transactions on Computer-Human Interaction* 29, 6 (Nov. 2022), 57:1–57:36.

https://doi.org/10.1145/3530013

[39] Chen-Yao Kao. 2020. How figurativity of analogy affects creativity: The application of four-term analogies to teaching for creativity. *Thinking skills and creativity* 36 (2020), 100653. Publisher: Elsevier.

[40] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[41] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3613904.3642773

[42] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting scientific results in perspective: Improving the communication of standardized effect sizes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 625, 14 pages. https://doi.org/10.1145/3491102.3502053

[43] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 38–48. https://doi.org/10.1145/2858036.2858440

[44] Varun Kumar, Savita Bhat, and Niranjan Pedanekar. 2015. Stickipedia: A search engine and repository for explanatory analogies. In *2015 IEEE 15th International Conference on Advanced Learning Technologies*. IEEE, 280–284.

[45] Ruijia Li, Yiting Wang, Chanjin Zheng, Yuan-Hao Jiang, and Bo Jiang. 2024. Generating Contextualized Mathematics Multiple-Choice Questions Utilizing Large Language Models. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 494–501.

[46] Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, and Jinjun Xiong. 2024. Large Language Models have Intrinsic Self-Correction Ability. *arXiv preprint arXiv:2406.15673* (2024).

[47] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 454, 18 pages. https://doi.org/10.1145/3544548.3580957

[48] Wenhan Lyu, Yimeng Wang, Tingting (Rachel) Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the Effectiveness of LLMs in Introductory Computer Science Education: A Semester-Long Field Study. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 63–74. https://doi.org/10.1145/3657604.3662036

[49] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 746–751. https://aclanthology.org/N13-1090

[50] Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences* 1505, 1 (2021), 79–101. Publisher: Wiley Online Library.

[51] Tricia J. Ngoon, S Sushil, Angela E.B. Stewart, Ung-Sang Lee, Saranya Venkatraman, Neil Thawani, Prasenjit Mitra, Sherice Clarke, John Zimmerman, and Amy Ogan. 2024. ClassInSight: Designing Conversation Support Tools to Visualize Classroom Discussion for Personalized Teacher Professional Development. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3642487

[52] Wode Ni, Sam Estep, Hwei-Shin Harriman, Kenneth R. Koedinger, and Joshua Sunshine. 2024. Edgeworth: Efficient and Scalable Authoring of Visual Thinking Activities. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 98–109. https://doi.org/10.1145/3657604.3662034

[53] José Oliva, Pilar Azcárate, and Antonio Navarrete Salvador. 2007. Teaching Models in the Use of Analogies as a Resource in the Science Classroom. *International Journal of Science Education - INT J SCI EDUC* 29 (Jan. 2007), 45–66. https://doi.org/10.1080/09500690600708444

[54] OpenAI. 2023. GPT-4 Technical Report. _eprint: 2303.08774.

[55] OpenAI. 2024. Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=TG8KACxEON

[57] Juho Pääkkönen, Matti Nelimarkka, Jesse Haapoja, and Airi Lampinen. 2020. Bureaucracy as a Lens for Analyzing and Designing Algorithmic Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376780

[58] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188* (2023).

[59] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[61] Lindsey Richland, Keith Holyoak, and James Stigler. 2004. Analogy Use in Eighth-Grade Mathematics Classrooms. *Cognition and Instruction* 22 (March 2004), 37–60. https://doi.org/10.1207/s1532690Xci2201_2

[62] Lindsey Engle Richland and Nina Simms. 2015. Analogy, higher order thinking, and education. *WIREs Cognitive Science* 6, 2 (2015), 177–192. https://doi.org/10.1002/wcs.1336 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1336.

[63] Lindsey E. Richland, Osnat Zur, and Keith J. Holyoak. 2007. Cognitive Supports for Analogies in the Mathematics Classroom. *Science* 316, 5828 (May 2007), 1128–1129. https://doi.org/10.1126/science.1142103 Publisher: American Association for the Advancement of Science.

[64] Christopher Riederer, Jake M. Hofman, and Daniel G. Goldstein. 2018. To Put That in Perspective: Generating Analogies that Make Numbers Easier to Understand. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3174122

[65] Jinxin Shi, Jiabao Zhao, Xingjiao Wu, Ruyi Xu, Yuan-Hao Jiang, and Liang He. 2025. Mitigating reasoning hallucination through Multi-agent Collaborative Filtering. *Expert Systems with Applications* 263 (2025), 125723. https://doi.org/10.1016/j.eswa.2024.125723

[66] Sofia Eleni Spatharioti, Daniel G Goldstein, and Jake M Hofman. 2024. Using Open Data to Automatically Generate Localized Analogies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642638 event-place: Honolulu, HI, USA.

[67] Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. ParallelPARC: A Scalable Pipeline for Generating Natural-Language Analogies. *arXiv preprint arXiv:2403.01139* (2024).

[68] Oren Sultan and Dafna Shahaf. 2022. Life is a Circus and We are the Clowns: Automatically Finding Analogies between Situations and Processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3547–3562. https://aclanthology.org/2022.emnlp-main.232

[69] Mei Tan and Hari Subramonyam. 2024. More than Model Documentation: Uncovering Teachers' Bespoke Information Needs for Informed Classroom Integration of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642592

[70] Xiaohang Tang, Sam Wong, Kevin Pu, Xi Chen, Yalong Yang, and Yan Chen. 2024. VizGroup: An AI-assisted Event-driven System for Collaborative Programming Learning Analytics. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 93, 22 pages. https://doi.org/10.1145/3654777.3676347

[71] Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. _eprint: 2312.11805.

[72] Paul Thagard. 1992. Analogy, explanation, and education. *Journal of research in science teaching* 29, 6 (1992), 537–544. Publisher: Wiley Online Library.

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[74] David Treagust, Reinders Duit, Paul Joslin, and Ivo Lindauer. 1992. Science teachers' use of analogies: Observations from classroom practice. *International Journal of Science Education - INT J SCI EDUC* 14 (Oct. 1992), 413–422. https://doi.org/10.1080/0950069920140404

[75] Peter D Turney, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. *Recent*

Advances in Natural Language Processing III: Selected Papers from RANLP 2003 (2003), 101–110.

[76] Michael S. Vendetti, Bryan J. Matlen, Lindsey E. Richland, and Silvia A. Bunge. 2015. Analogical Reasoning in the Classroom: Insights From Cognitive Science. *Mind, Brain, and Education* 9, 2 (2015), 100–106. https://doi.org/10.1111/mbe.12080 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mbe.12080.

[77] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2024. ReelFramer: Human-AI Co-Creation for News-to-Video Translation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 169, 20 pages. https://doi.org/10.1145/3613904.3642868

[78] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2022. Emergent Analogical Reasoning in Large Language Models. *arXiv preprint arXiv:2212.09196* (2022).

[79] Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 3534–3549. https://doi.org/10.18653/v1/2023.findings-acl.218

[80] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112.

[81] ShunYi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault. 2024. Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 806, 32 pages. https://doi.org/10.1145/3613904.3642530

[82] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1245–1254. https://doi.org/10.1145/2556288.2557371

[83] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Searching for analogical ideas with crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1225–1234. https://doi.org/10.1145/2556288.2557378

[84] Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling Script Knowledge from Large Language Models for Constrained Language Planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4303–4325. https://doi.org/10.18653/v1/2023.acl-long.236

[85] Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023. Beneath Surface Similarity: Large Language Models Make Reasonable Scientific Analogies after Structure Abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2446–2460. https://doi.org/10.18653/v1/2023.findings-emnlp.160

[86] Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023. ANALOGYKB: Unlocking Analogical Reasoning of Language Models with A Million-scale Knowledge Base. *arXiv preprint arXiv:2305.05994* (2023).

[87] Ashley Ge Zhang, Xiaohang Tang, Steve Oney, and Yan Chen. 2024. CFlow: Supporting Semantic Flow Analysis of Students' Code in Programming Problems at Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 188–199. https://doi.org/10.1145/3657604.3662025

[88] Qihao Zhu, Xinyu Zhang, and Jianxi Luo. 2023. Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. *Journal of Mechanical Design* 145, 4 (01 2023), 041409. https://doi.org/10.1115/1.4056598 arXiv:https://asmedigitalcollection.asme.org/mechanicaldesign/article-pdf/145/4/041409/6974748/md_145_4_041409.pdf